Test Fairness and Validity of the TEPS

Inn Chull Choi

A great deal of public attention has been given to the Test of English Proficiency, developed by Seoul National University (TEPS), which has been administered as an official test of general English proficiency since January 1999. In order for TEPS to serve as a valid and reliable measurement tool, it is imperative that test developers consider the notion of test fairness including the washback (backwash) effect (or impact: Bachman & Palmer 1996) as one of the most fundamental considerations. Therefore, the present study is intended to validate TEPS, based largely on test-takers' qualitative and quantitative feedback on the pilot TEPS and the first administrated TEPS. The findings from the survey and the statistical analyses reveal that TEPS has fulfilled the basic requirement of validity with adequate test fairness by satisfying the high expectation of test-takers and test-users.

I. Validity of TEPS as an Indirect Measurement Tool of Communicative Competence

The objective test by definition makes it impossible to elicit the test-taker's speech sample. Thus, the currently available so-called objective test format is inherently incapable of measuring the test-taker's overall oral proficiency or his or her productive skills. With serious consideration of valid test method facets based on sound theories of language acquisition and use, however, it may be possible to indirectly measure the test-taker's overall proficiency or productive skills. As Choi (1997) indicated, TEPS employs some specially designed test methods. Among the unique features different from the conventional formats are 1) exposure to oral input only, 2) two-time exposure (macro-listening and micro-listening) for listening comprehension tests, 3) one passage one item (OPOI) for listening comprehension and reading comprehension tests, 4) maximized speededness

and 5) separate consideration of written style and spoken style English for vocabulary and grammar tests.

The validity of a proficiency test can be investigated by predicting a test-taker's performance in a real-world communicative setting on the basis of his or her test performance. In reality, however, it is seldom logistically feasible to conduct such a validation exercise. Therefore, an alternative way of validation is to explore the extent to which the outcome of the test in question correlates concurrently with the performance based on a valid speaking test or oral proficiency interview. For the purpose of the present validation study, the Test of Oral Proficiency or TOP is employed to serve as a valid tool to measure oral proficiency. TOP is a simulated oral proficiency interview (SOPI: Stansfield & Kenyon, 1992) test developed by Seoul National University Language Research Institute. TOP has been validated through rigorous quantitative and qualitative analyses (최인철 1998). Among 310 Seoul National University students who took the experimental version of TEPS on December 5, 1997, 56 subjects voluntarily participated in this study to take TOP.

The findings from the correlational study on TEPS and TOP are presented as follows.

1. Correlations between TEPS & TOP (Test of Oral Proficiency)

1.1. Overall High Correlations between TEPS and TOP

Table 1 shows that there are significantly high correlation coefficients between TEPS and TOP across the components of each test except for only a few components in reading and vocabulary. This indicates that valid methods of an objective test (measuring receptive skills) can measure oral proficiency indirectly fairly well. This finding also concurs with our own experience that receptive skills are somewhat independent of productive skills, but are essentially intertwined with productive skills.

1.2. Very High Correlations between TEPS LC and TOP

The table also demonstrates the high correlation coefficients between TEPS listening comprehension test results and the five components of TOP (i.e, pronunciation, grammar, fluency, overall comprehensibility, appropriateness) are higher than .6. These coefficients are as high as those between TOEFL I (listening comprehension) and FSI and TSE (Clark & Swinton

Table 1. TOP-TEPS Correlational Coefficients.

	LC	RC	GR	VC	TOT	PRON	GRAM	FLUN	COMP	APPR
LC	1.0000	.5025**	.6307**	.4239*	.8932**	.6013**	.7001**	.7247**	.7048**	.6623**
RC	.5025**	1.0000	.5111**	.5189**	.8100**	.2438	.4340**	.3796*	.3757*	.3938*
GR	.6307**	.5111**	1.0000	.6805**	.7623**	.4152*	.5258**	.5576**	.5271**	.5142**
VC	.4239*	.5189**	.6805**	1.0000	.6516**	.2400	.3764*	.3914*	.3548*	.3507*
TOT	.8932**	.8100**	.7623**	.6516**	1.0000	.5159**	.6780**	.6755**	.6546**	.6359**
PRON	.6013**	.2438	.4152*	.2400	.5159**	1.0000	.8770**	.9260**	.9264**	.9038**
GRAM	.7001**	.4340**	.5258**	.3764*	.6780**	.8770**	1.0000	.9637**	.9676**	.9349**
FLUN	.7247**	.3796*	.5576**	.3914*	.6755**	.9260**	.9637**	1.0000	.9735**	.9588**
COMP	.7048**	.3757*	.5271**	.3548*	.6546**	.9264**	.9676**	.9735**	1.0000	.9578**
APPR	.6623**	.3938*	.5142**	.3507*	.6359**	.9038**	.9349**	.9588**	.9578**	1.0000

N of cases : 56 2-tailed Signif : * - .01 ** - .001
" . " is printed if a coefficient cannot be computed
LC : TEPS Listening Comprehension
RC : TEPS Reading Comprehension
GR : TEPS Grammar
VC : TEPS Vocabulary
TOT : TEPS Total Score
PRON : TOP Pronunciation
GRAM : TOP Grammar
FLUN : TOP Fluency
COMP : TOP Comprehensibility
APPR : TOP Appropriateness

1980). This reveals that the listening comprehension test of TEPS is a highly valid indirect measurement tool to assess overall communicative competence. This finding is endorsed by the well-documented fact that listening skills, constituting an essential core of overall proficiency, have greater transitional effects on communicative competence than any other skill. Thus, effective language courses following the natural approach have put greater emphasis on the receptive listening skills than the productive oral skills.

1.3. Relatively High Correlations between TEPS GR and TOP

The table also illustrates fairly high correlation coefficients between the TEPS grammar test and the four components of TOP (excluding pronunciation) which are higher than .5. This finding reveals that the grammar test of TEPS is a valid measurement tool to assess the acquisition (Krashen 1985) or the subconscious communicative competence, which can be activated for oral proficiency.

These positive findings can be attributed to the valid test method facets

of the TEPS grammar test such as 1) maximized speededness (which inhibits the use of learning and activates the acquisition), 2) separate measurement of spoken style language usage as well as written style language usage, 3) enhanced context-embeddedness through providing contextualized stems excluding the error detection task type with underlined format.

1.4. Low Correlations between TEPS RC and VC and TOP

The table shows the fairly low correlation coefficients between the reading comprehension and listening comprehension of TEPS and the five components of TOP. This finding is to be expected since TOP is designed to measure the ability to use spoken language, whereas the reading comprehension test and the written style content of the vocabulary test of TEPS have contents designed solely to reflect the use of written language. This finding supports the discriminant or divergence (Bachman & Eignor 1997) validity of TEPS reading comprehension and vocabulary tests.

1.5. High Correlation between TEPS Total Scores and TOP Scores

Finally, the table reveals that all the correlation coefficients between the total score of TEPS and the five components of TOP (excluding pronunciation: .5159) are higher than .63. This finding constitutes convincing evidence that TEPS is a valid measurement tool to assess communicative competence in an indirect manner.

2. Measurement of Communicative Competence as Demonstrated by TEPS Results & Survey

The final pilot test of TEPS was administered to 310 SNU students on December 5, 1997. A case study or qualitative research was conducted on the test takers whose TEPS ability levels were higher than level 1 and whose TOEIC (Test of English for International Communication) scores were available. Noteworthy are the following findings.

2.1. Validity of TEPS in Assessing Communicative Competence

The majority of students with TEPS ability higher than level 1 had the

experience of living in English speaking countries before they reached the age of puberty. Thus they had acquired English in a natural way or a real immersion method and achieved the level of near-native English competence. This finding is congruent with that of the aforementioned correlational study.

2.2. Limited Comparability between TEPS & TOEIC Results

As it was virtually impossible to have the subjects take both TEPS and TOEIC concurrently and obtain their item-level responses, a quantitative research method cannot be employed to conduct a comparability study. Therefore, it was decided that an interview would be conducted with those test-takers who said through the survey that they had been exposed to an English speaking environment during the period of their childhood.

2.2.1. TEPS score Lower than TOEIC score

Table 2 and the continuing table presented in Appendix 1. show that almost all the cases in the present research fell into this category. According to the level description of TOEIC, those test takers with scores higher than 860 are considered to have adequate communicative competence as a non-native speaker. In this present study, this claim was not necessarily the case, as was revealed by the interview with the subjects who had more than 860 but admitted that they did not have adequate communicative competence or found it difficult to make themselves understood with ease.

Among the test-takers with TEPS level 1+ to 2+, there was a tendency for a lower TEPS score than TOEIC score. Around TEPS level 3, no obvious pattern seemed to exist, i.e., there were those with TEPS scores higher than TOEIC, and vice versa.

2.2.2. TEPS score Higher than TOEIC score (shown in shade)

Table 2 also reveals that only a few cases showed that the test-takers' TEPS total scores were higher than TOEIC total scores.

There was a test-taker who lived in Canada for 36 months (from the age of seven till the age of nine.) In spite of his native-like English competence, he obtained only 880 on TOEIC. On the other hand, he achieved the level 1+ on TEPS, which is described as being equivalent to the native level of communicative competence. The same was true of his elder brother and

elder sister (a professional English instructor) who have a good command of English. They obtained 885, 920 respectively on TOEIC. It was the first time for them to take both TOEIC and TEPS, which suggests that the effect of the factor of test-taking strategy or 'test wiseness' on test scores was minimized. Thus, it can be safely claimed that their test results manifest a very accurate measurement of communicative competence. This finding reveals that TEPS can constitute a more valid measurement tool of communicative competence than other systems currently available.

2.2.3. Validity of Listening Comprehension Test of TEPS

It is also worth noting that all of those test-takers who had been exposed to English in English speaking countries for more than one year before they reached the age of puberty succeeded in achieving the level 1+ on TEPS.

On the other hand, none of those test-takers who lived in English speaking countries for about one year to learn English after entering college, managed to achieve the level 1+ or 1 on TEPS. This finding strongly endorses the Critical Period Hypothesis (Lenneberg 1967). It also suggests that the overseas English learning program for college students may not prove to be as fruitful as expected.

In the following table, the country and the following numbers (right to the level of total score) indicate the English-speaking country the test takers stayed in and the total number of months they were exposed to English in that particular country. The two numbers within the parentheses represent their starting and ending age. Only one number representing the age is given in cases when the respondent stayed in the country less than one year.

Table 2. TEST Results of TEPS & Communicative Competence/ Demographic Background.

Rk	LC	Lv	RC	Lv	GR	Lv	VC	Lv	Tot	Lv	Self-rated Proficiency/Demographic Background/Comment
1	387	1+	380	1+	92	1+	94	1+	953	1+	US 9(0-1); TOEFL 640; advanced oral skills
2	367	1+	390	1+	92	1+	86	1	935	1+	None; began learning English from 3rd grade; listen to AFKN; advanced proficiency
3	360	1	390	1+	90	1	82	1	922	1+	None; began learning English from kindergarten; TOEIC 975; advanced proficiency
4	380	1+	350	1	94	1+	96	1+	920	1+	US 7(7-12); TOEFL 650; near-native proficiency
5	380	1+	370	1+	82	1	74	2+	906	1+	Singapore 48(7-9); advanced proficiency
6	367	1+	380	1+	82	1	74	2+	903	1+	None; began learning English from the elementary school; TOEFL 600; advanced proficiency

Rk	LC	Lv	RC	Lv	GR	Lv	VC	Lv	Tot	Lv	Self-rated Proficiency/Demographic Background/Comment
78	387	1+	360	1	82	1	72	2+	901	1+	Canada 36(7-9); TOEIC 880 IELTS 7/9; advanced proficiency; (brother) 36(9-11) TOEIC 885; (sister) 36(10-13) TOEFL 647, TOEIC 920 advanced proficiency
8	353	1	380	1+	84	1	80	2+	897	1	None; OPOI is better than OPMI
9	360	1	370	1+	82	1	84	1	896	1	None; TOEIC 895; served military duty in KATUSA; advanced proficiency; LC test method reduces unwarranted memory load and enhances concentration
10	367	1+	360	1	84	1	78	2+	889	1	US 26(10-12); TOEFL 593; advanced proficiency
11	367	1+	360	1	86	1	76	2+	889	1	UK 40(4-7); advanced proficiency
12	353	1	360	1	82	1	92	1+	887	1	None; KATUSA; TOEIC 940; advanced proficiency
13	373	1+	340	1	88	1	84	1	885	1	Canada; 36(9-12); advanced proficiency
14	380	1+	350	1	74	2+	80	2+	884	1	US; 11(8-9); TOEIC 920 TOEFL 600; advanced proficiency
15	367	1+	340	1	84	1	92	1+	883	1	US; 2(14-15); TOEFL 627
16	380	1+	330	1	82	1	90	1	882	1	None; TOEFL 647; advanced proficiency; desirable Vocabulary and RC tests methods
17	387	1+	340	1	80	2+	70	2	877	1	US; 11(12); advanced proficiency
18	353	1	360	1	76	2+	88	1	877	1	None; TOEFL 623; desirable LC's 'oral input only' method
19	340	1	380	1+	84	1	72	2+	876	1	None; desirable LC test method
20	367	1+	340	1	82	1	86	1	875	1	Canada; 9(19)
21	333	1	370	1+	84	1	86	1	873	1	US; 5(22); Overall, RC is difficult
22	367	1+	330	1	84	1	92	1+	873	1	None; TOEFL 620; a variety of topics
23	353	1	350	1	74	2+	90	1	867	1	US; 12(20-21)
24	333	1	360	1	88	1	84	1	865	1	None;
25	333	1	380	1+	72	2+	78	2+	863	1	None; TOEIC 850; speededness improves discrimination
26	360	1	340	1	82	1	80	2+	862	1	US; 10(14); TOEFL 615 TOEIC 930
27	367	1+	330	1	90	1	74	2+	861	1	Hong Kong; 30(8-10); advanced proficiency
28	367	1+	340	1	74	2+	78	2+	859	1	None;
29	367	1+	340	1	80	2+	72	2+	859	1	None; high intermediate proficiency
30	333	1	370	1+	78	2+	74	2+	855	1	None; desirable test methods
31	320	2+	380	1+	78	2+	76	2+	854	1	No response
32	320	2+	380	1+	64	2	90	1	854	1	No response
33	333	1	370	1+	66	2	82	1	851	1	No response
34	360	1	320	2+	82	1	84	1	846	1	None; LC is as difficult as TOEIC; RC is more difficult than TOEIC
35	373	1+	300	2+	90	1	80	2+	843	1	US; 22(10-12); practical English oriented content
36	353	1	320	2+	80	2+	82	1	835	1	No response
37	340	1	330	1	78	2+	86	1	834	1	No response
38	353	1	310	2+	84	1	86	1	833	1	None; desirable test methods
39	347	1	340	1	78	2+	68	2	833	1	No response
40	367	1+	320	2+	70	2	74	2+	831	1	No response
41	313	2+	380	1+	68	2	68	2	829	1	No response
42	327	1	360	1	68	2	74	2+	829	1	US; 10(21)
43	320	2+	350	1	80	2+	78	2+	828	1	No response
44	320	2+	350	1	78	2+	80	2+	828	1	US; 2(25); TOEIC 940; intermediate proficiency; grammar test is difficult
45	340	1	330	1	76	2+	78	2+	824	1	None; TOEFL 580
46	353	1	330	1	66	2	72	2+	821	1	None; TOEFL 633; more difficult than TOEFL, TOEIC
47	347	1	300	2+	88	1	84	1	819	1	No response
48	333	1	330	1	78	2+	76	2+	817	1	None; TOEFL 623
49	313	2+	340	1	76	2+	88	1	817	1	No response
50	340	1	330	1	68	2	78	2+	816	1	US; 1(19); highly valid test content

* The remaining content of the table is presented in Appendix 1.
* LC: Listening Comprehension; GR: Grammar; VC: Vocabulary; RC: Reading Comprehension; Tot: Total
* Rk: rank; Lv: level

Ⅱ. Descriptive Statistics of the First TEPS

The following is the descriptive statistics of TEPS first administered to 4569 test-takers on January 31, 1999. It should be noted that the following statistics based on the classical testing theory are to be used for reference, in that TEPS was developed and analysed within the theoretical framework of the Item Response Theory. The indices of skewness and kurtosis of the four subtests vary within the range +1 and -1, which is much narrower than the rule of thumb criterion of normal distribution (+2 and -2), thus showing that overall test performance does not violate the normal distribution assumption.

1. Listening Comprehension

Table 3. Descriptive Statistics.

N of Items	60
N of Examinees	4569
Mean	39.121
Variance	102.532
Std. Dev.	10.126
Skew	-0.267
Kurtosis	-0.530
Minimum	0.000
Maximum	60.000
Median	40.000
Alpha	0.907
SEM	3.089
Mean P	0.652
Mean Item-Tot.	0.398
Mean Biserial	0.563
Max Score (Low)	33
N (Low Group)	1339
Min Score (High)	46
N (High Group)	1358

According to Table 3, the Cronbach alpha is .907, which shows the listening comprehension test proves to be quite reliable from the perspective of classical testing theory. The standard error of measurement (SEM) is 3.089 out of 60 points. For our rescaled test score report, the maximum possible score for the listening comprehension is set at 400. Hence, the

extrapolated SEM is approximately 20.60.

The overall discrimination power index manifested by the mean item-total correlation is almost .4, which is over the adequate criterion of .3, thus showing that the test succeeds in discriminating among the test-takers of the total test-taker group fairly well. This high degree of discrimination can be clearly supported by the negative index of kurtosis and the dispersion among the upper group and the lower group — the maximum score of the low ability level group is 33, whereas the minimum score of the high ability level group is 46.

The difficulty/facility index of mean P (proportion correct) is .652, which is slightly over the appropriate range of .5 to .6. The slightly negative index of skewness and the item analysis (as in Appendix 2) suggest that the listening comprehension test proves to be a bit easy, especially for the high ability group. This finding can be accounted for by the demographic survey which shows that this first test was taken by the advanced level group (higher than the normal target group) including many English lecturers of private English institutes, where spoken English is more emphasized than written English.

2. Grammar

Table 4. Descriptive Statistics.

N of Items	50
N of Examinees	4569
Mean	27.355
Variance	76.439
Std. Dev.	8.743
Skew	0.063
Kurtosis	-0.462
Alpha	0.878
SEM	3.052
Mean P	0.547
Mean Item-Tot.	0.380
Mean Biserial	0.503
Max Score (Low)	22
N (Low Group)	1370
Min Score (High)	33
N (High Group)	1323

According to Table 4, the Cronbach alpha is .878, which shows the grammar test is quite reliable from the perspective of the classical testing theory. The standard error of measurement (SEM) is 3.052 out of 40 points. For our rescaled test score report, the maximum possible score for the grammar test is set at 100. Hence, the extrapolated SEM is approximately 7.63.

The overall discrimination power index or the mean item-total correlation is .38, which is over the adequate criterion of .3, thus showing that the test succeeds in discriminating among the test-takers in the total group fairly well. This high degree of discrimination can be clearly supported by the negative index of kurtosis and the dispersion among the upper group and the lower group — the maximum score of the low ability level group is 22, whereas the minimum score of the high ability level group is 33.

The difficulty/facility index of mean P is .547, which is within the appropriate range of .5 to .6, as is also shown by the near zero index of skewness.

3. Vocabulary

Table 5. Descriptive Statistics.

N of Items	50
N of Examinees	4569
Mean	27.980
Variance	78.067
Std. Dev.	8.836
Skew	0.030
Kurtosis	-0.605
Alpha	0.885
SEM	2.995
Mean P	0.560
Mean Item-Tot.	0.386
Mean Biserial	0.514
Max Score (Low)	22
N (Low Group)	1345
Min Score (High)	34
N (High Group)	1330

According to Table 5, the Cronbach alpha is .885, which indicates the test is quite reliable from the standpoint of classical testing theory. The

standard error of measurement (SEM) is 2.995. For our rescaled test score report, the maximum possible score for the vocabulary is set at 100. Hence, the extrapolated SEM is approximately 7.49.

The overall discrimination power index manifested by the mean item-total correlation is .386, which is over the adequate criterion of .3, thus indicating that the test succeeds in discriminating the total test-taker group fairly well. This high degree of discrimination can be clearly supported by the negative index of kurtosis and the dispersion among the upper group and the lower group — the maximum score of the low ability level group is 22, whereas the minimum score of the high ability level group is 34.

The difficulty/facility index of mean P is .560, which is within the appropriate range of .5 to .6, as is also shown by the near zero index of skewness.

4. Reading Comprehension

Table 6. Descriptive Statistics

N of Items	40
N of Examinees	4569
Mean	23.682
Variance	52.840
Std. Dev.	7.269
Skew	-0.197
Kurtosis	-0.464
Alpha	0.861
SEM	2.713
Mean P	0.592
Mean Item-Tot.	0.396
Mean Biserial	0.526
Max Score (Low)	19
N (Low Group)	1315
Min Score (High)	29
N (High Group)	1260

According to Table 6, the Cronbach alpha is .861, which shows the test proves to be quite reliable from the standpoint of the classical testing theory. The standard error of measurement (SEM) is 2.713 out of 40 points. For our rescaled test score report, the maximum possible score for the reading comprehension is set at 400. Hence, the extrapolated SEM is

approximately 27.13

The overall discrimination power index manifested by the mean item-total correlation is almost .4, which is over the adequate criterion of .3 and shows that the test succeeds in discriminating the total test-taker group fairly well. This high degree of discrimination can be clearly supported by the negative index of kurtosis and the dispersion among the upper group and the lower group — the maximum score of the low ability level group is 19, where as the minimum score of the high ability level group is 29.

The difficulty/facility index of mean P is .592, which is within the appropriate range of .5 to .6, as is also shown by the near zero index of skewness.

Ⅲ. Analysis of IRT-based Test Results of 1st TEPS

1. Dimensionality Check

In order for IRT to be employed to analyze test results, it is important for the strong assumption of 'essential unidimensionality' to be met. The most popular way of checking dimensionality is Stout's method to investigate the extent to which a test is essentially unidimensional. It has been proposed as a nonparametric model to produce a statistical index, T, which indicates the extent of departure from unidimensionality (Stout et al. 1991). Exploratory factor analysis of tetrachoric inter-item correlation matrices was employed to create assessment subtests for the present study. The summary results are provided in Table 7, below. The eigenvalues from the factor analysis for each subtest are presented in Appendix 2.

NB: Under Ho: dimensionality = 1, T should be N(0,S), S < 1
Under H1: dimensionality > 1, T will have a positive mean.

Table 7. Summary Results of Stout's Approach.

Test	T	p	Ho (alpha = .01)	Dimensionality
Listening Comprehension	-6.839828	1.000000	accept	unidimensional
Grammar	-2.515826	.994062	accept	unidimensional
Vocabulary	1.018371	.154251	accept	unidimensional
Reading Comprehension	-1.711156	.956474	accept	unidimensional

For the listening comprehension test, the probability of the test essentially being unidimensional is 1.0, or the probability of the test necessarily being multidimensional is almost zero. For the grammar test, the probability of the test essentially being unidimensional is .9946062, or the probability of the test necessarily being multidimensional is almost zero (= 1.0-.994062). For the vocabulary test, the probability of the test essentially being unidimensional is .154251, or the probability of the test necessarily being multidimensional is .845749 (=1.0-.154251). Thus, it can be safely claimed that the vocabulary test content does not violate the essential unidimensionality assumption. It is impossible to reject the hypothesis of essential unidimensionality at the alpha level of .01. For the reading comprehension test, the probability of the test essentially being unidimensional is .956474, or the probability of the test necessarily being multidimensional is .043526 (=1.0-.956474). In conclusion, the test batteries are all proven to be essentially unidimensional, which validates the application of the item response theory to our study.

2. Comparison between CTT Observed Scores and IRT True Scores

2.1. Comparison between CTT Composite Scores and CTT Raw Scores

Table 8 shows the comparison between CTT scores and IRT true scores of the top 50 test takers. It illustrates that there was a significant discrepancy between the raw score (the total number of items correct) and the composite score based on varying weights on different subtests, i.e. four points for each item in the reading and listening comprehension test and one point for each item in the grammar and vocabulary tests. This clearly shows that it would be problematic or cause biased test results to simply calculate the total raw score by adding up the number of items correct without giving varying weights on different skill subtests. It would be reasonable to consider the relative complexity of cognitive processes required for solving the test items and the testing time in determining the varying weights.

Table 8. Comparison between CTT Scores and IRT Scores.

LC CTT	LC IRT	L	GR CTT	GR IRT	L	VC CTT	VC IRT	L	RC CTT	RC IRT	L	ALL CTT	ALL Rank	ALL Comp	ALL Rank	ALL IRT	L	ALL Rank
60	374	1+	49	88	1	49	92	1+	39	374	1+	197	1	494	1	928	1+	4
59	374	1+	48	88	1	49	92	1+	40	374	1+	196	2	493	2	928	1+	1

LC CTT	LC IRT	L	GR CTT	GR IRT	L	VC CTT	VC IRT	L	RC CTT	RC IRT	L	ALL CTT	Rank	ALL Comp	Rank	ALL IRT	L	Rank
60	374	1+	50	88	1	48	92	1+	38	368	1+	196	3	490	3	922	1+	14
58	374	1+	50	88	1	47	92	1+	40	374	1+	195	5	489	5	928	1+	3
58	374	1+	48	88	1	49	92	1+	40	374	1+	195	4	489	4	928	1+	2
57	362	1+	49	88	1	49	92	1+	40	374	1+	195	6	486	6	916	1+	22
58	372	1+	49	88	1	47	91	1+	39	374	1+	193	7	484	10	926	1+	7
58	374	1+	47	88	1	49	92	1+	39	374	1+	193	9	484	9	928	1+	5
60	374	1+	49	88	1	47	89	1	37	357	1	193	8	484	11	908	1+	37
58	374	1+	49	88	1	48	92	1+	38	366	1+	193	10	481	14	920	1+	18
58	374	1+	48	88	1	45	88	1	40	374	1+	191	11	485	8	924	1+	9
58	371	1+	49	88	1	46	89	1	38	372	1+	191	12	479	18	920	1+	17
59	374	1+	46	87	1	47	92	1+	38	374	1+	190	13	481	15	927	1+	6
58	374	1+	50	88	1	44	87	1	38	365	1+	190	15	478	20	914	1+	27
59	374	1+	48	88	1	46	92	1+	37	352	1	190	14	478	21	916	1+	24
55	368	1+	48	88	1	48	92	1+	39	374	1+	190	16	472	29	922	1+	13
59	374	1+	45	85	1	45	90	1	40	374	1+	189	17	486	7	923	1+	10
59	374	1+	47	88	1	44	86	1	39	374	1+	189	18	483	12	922	1+	15
60	374	1+	42	77	2+	49	92	1+	38	374	1+	189	19	483	13	917	1+	19
59	374	1+	48	88	1	44	89	1	38	374	1+	189	20	480	16	925	1+	8
60	374	1+	47	85	1	46	90	1	36	348	1	189	21	477	24	897	1	60
57	360	1	47	88	1	47	92	1+	38	371	1+	189	22	474	27	911	1+	32
57	356	1	49	88	1	46	91	1+	37	368	1+	189	23	471	30	903	1+	43
56	364	1+	47	88	1	46	90	1	39	374	1+	188	24	473	28	916	1+	23
58	374	1+	46	88	1	48	92	1+	36	361	1+	188	26	470	34	915	1+	26
59	374	1+	49	88	1	45	90	1	35	357	1	188	25	470	33	909	1+	35
57	362	1+	47	88	1	47	92	1+	37	356	1	188	27	470	35	898	1	53
56	360	1	46	87	1	49	92	1+	37	373	1+	188	28	467	40	912	1+	29
60	374	1+	50	88	1	46	87	1	32	317	2+	188	29	464	44	866	1	47
54	352	1	49	88	1	48	92	1+	37	366	1+	188	30	461	46	898	1	52
59	374	1+	47	88	1	43	87	1	38	374	1+	187	31	478	19	923	1+	12
57	368	1+	45	87	1	48	92	1+	37	354	1	187	33	469	36	901	1+	48
56	359	1	46	86	1	47	91	1+	38	374	1+	187	32	469	37	910	1+	33
58	370	1+	47	88	1	47	92	1+	35	348	1	187	34	466	42	898	1	51
55	353	1	48	88	1	46	90	1	38	374	1+	187	35	466	43	905	1+	38
59	374	1+	45	85	1	43	84	1	39	373	1+	186	36	480	17	916	1+	25
58	374	1+	43	85	1	46	90	1	39	374	1+	186	38	477	22	923	1+	11
59	374	1+	43	83	1	46	88	1	38	366	1+	186	37	477	23	911	1+	31
59	374	1+	46	88	1	45	89	1	36	357	1	186	39	471	31	908	1+	36
58	371	1+	48	88	1	44	88	1	36	351	1	186	41	468	38	898	1	54
59	374	1+	46	88	1	46	90	1	35	343	1	186	40	468	39	895	1	64
58	374	1+	46	87	1	48	92	1+	34	329	1	186	42	462	45	882	1	96
60	374	1+	47	88	1	41	79	2+	37	361	1+	185	44	476	25	902	1+	46
58	374	1+	44	84	1	44	88	1	39	374	1+	185	43	476	26	920	1+	16
57	364	1+	47	88	1	43	84	1	38	374	1+	185	45	470	32	910	1+	34
58	374	1+	47	88	1	44	87	1	36	356	1	185	46	467	41	905	1+	41
58	373	1+	47	88	1	47	92	1+	33	330	1	185	47	458	47	883	1	92
54	358	1	48	88	1	46	89	1	37	367	1+	185	48	458	48	902	1+	44
56	362	1+	49	88	1	46	90	1	34	353	1	185	49	455	49	893	1	68

* L represents Ability Level; Comp represents composite score

2.2. Comparison between CTT Raw Scores and IRT True Scores

The scores of the four test takers with the identical CTT raw scores (193) are extracted from Table 8 to be given in Table 9. The table highlights the fact that there can be a significant discrepancy between CTT raw scores and IRT true scores. The table demonstrates that those test takers with identical CTT Scores (193) and thus the same rank (7th) proved to have different IRT true scores (908 to 928) and thus the varying ranks (from 7th to 37th). The table, based on these particular IRT true scoring results, also indicates that the items of the reading comprehension test carry more weight than those of the listening comprehension test. That is, the test-takers who got 37 items correct on the reading comprehension test got an IRT true score of 357, and those who got 39 items correct got an IRT true score of 374. On the other hand, the test-takers who got 58 items correct on the listening comprehension test got an IRT true score of 374 and those who got 60 items correct got the same IRT true score of 374. It also suggests that the items of the grammar test and the vocabulary test carry much less discriminating power. That is, the test-takers who got 47 items correct on the grammar test got an IRT true score of 88 and those who got 49 items got the same IRT true score. The test-takers who got 47 items correct on the vocabulary test got an IRT true score of 89 or 91 and those who got 49 items correct got an IRT true score of 92.

Table 9. Comparison between CTT Scores and IRT Scores of Four Test-takers with Identical Rank based on CTT Scores and Different Ranks based on IRT Scores.

LC CTT	LC IRT	L	GR CTT	GR IRT	L	VC CTT	VC IRT	L	RC CTT	RC IRT	L	ALL CTT	Rank	ALL Comp	Rank	ALL IRT	L	Rank
58	372	1+	49	88	1	47	91	1+	39	374	1+	193	7	484	10	926	1+	7
58	374	1+	47	88	1	49	92	1+	39	374	1+	193	7	484	9	928	1+	5
60	374	1+	49	88	1	47	89	1	37	357	1	193	7	484	11	908	1+	37
58	374	1+	49	88	1	48	92	1+	38	366	1+	193	7	481	14	920	1+	18

* L denotes Ability Level

2.3. Distribution of Numbers of Test-takers across Levels

As Table 10 suggests, the ability distribution of the test-taker group reflects the normal distribution. There were more test takers in level 2 than

other ability levels. The table also shows that there was an almost zero percent of test-takers in the level 5 and 5+. This clearly indicates that the ability in general of the test-taker group is above average.

The table reveals that the grammar and the vocabulary test is a bit difficult for the majority of test-takers. It also indicates that the majority of students find the listening comprehension test a bit easier than the reading comprehension test. Witnessed repeatedly including in the pilot test results, this phenomenon may have to do with the current trend in which more emphasis has been placed on oral communication skills (listening and speaking) rather than written communication skills (reading and writing). In order for the TEPS to have a positive washback effect on English education, it is considered desirable to strike a balance in the current levels of difficulty between listening comprehension (spoken style English) and reading comprehension (written style English).

Table 10. Distribution of Number of Test-takers across the Different Tests and Levels.

Tests / Level	Listening Comp		Reading Comp		Grammar		Vocabulary		Total Score	
	No. of test-takers	(%)	No. of test-takers	(%)	No. of test-takers	(%)	No. of test-takers	(%)	No. of test-takers	(%)
1+	177	4	102	2	0	0	57	1	48	1
1	770	17	491	11	231	5	354	8	489	11
2+	1035	23	790	17	515	11	687	15	876	19
2	934	20	960	21	892	20	866	19	1046	23
3+	767	17	876	19	984	22	860	19	976	21
3	495	11	661	14	939	21	837	18	665	15
4+	291	6	464	10	584	13	612	13	368	8
4	93	2	189	4	367	8	262	6	96	2
5+	5	0	22	0	51	1	24	1	3	0
5	2	0	14	0	6	0	10	0	2	0

IV. Validity of Test Methods as Perceived by Test-takers

Questionnaires in language test validation have been used to capture test-takers' reactions to and opinions of the quality of test items and the 'face validity' of the test methods (Bradshaw 1990, Brown 1997). Given the fact that the test method facets in question are based on a sound theoretical framework of language testing (Bachman 1996, Oller 1995, Choi 1997), it may well be worthwhile to explore the degree of validity perceived

by the test-takers and the potential test-users who actually have scrutinized the test method facets through taking the test.

1. Survey on the Pilot Test

314 Seoul National University (SNU) students were invited to participate in a survey after taking the pilot TEPS. The following is a summary of the survey results.

NB: Value 1: Strongly disagree
Value 2: Somewhat disagree
Value 3: Neutral response
Value 4: Somewhat agree
Value 5: Strongly agree

Questionnaire items 1-6 concern the test-takers' self-rating of their listening comprehension (LC), grammatical competence (GC), vocabulary power (VP), reading comprehension (RC), and conversation skills (CS). The self-rating and some other results are not provided here as they are not directly related to the objective of the present study.

Table 11. LC: Validity of Double Exposure to Listening Passage & Question.

Value	Frequency	Percent
1	31	9.9
2	43	13.7
3	9	2.9
4	92	29.3
5	139	44.3
Mean		3.84

As for the listening comprehension test, it is deemed considerably valid for the test-takers to go through the test-taking process required by the test format, i.e., to be exposed twice to a listening passage and a question before listening to four choices. This clearly indicates that the majority of test-takers are in favor of this test method, which is designed to reflect the natural cognitive processes (macro-listening followed by micro-listening) required for successful listening activity in real-world communication

settings. The survey shows that most test-takers think highly of this test method, whereas those test-takers with near-native level of English proficiency do not appear to endorse this test method, especially in Part 1 and Part 2 (task of choosing the appropriate response).

Table 12. LC: Test Content Presented Only in Aural Mode.

Value	Frequency	Percent
1	17	5.4
2	55	17.5
3	11	3.5
4	97	30.9
5	134	42.7
Mean		3.88

As was mentioned previously, a traditional method for testing listening comprehension skills results in an adulterated score combining the ability to read and to listen and is therefore not a valid measurement tool of listening skills. The respondents appear to understand that there are some inherent problems with this kind of testing — seriously negative effects on the construct validity and interactiveness of the test. The above table demonstrates that many respondents are in support of the test method of presenting the listening material only in aural mode. The analysis of the respondents' feedback reveals that this favorable reaction is especially true of the high level ability test-takers.

Table 13. LC: OPOI Reduces Unwarranted Memory Load.

Value	Frequency	Percent
1	4	1.3
2	29	9.2
3	8	2.5
4	112	35.7
5	161	51.3
Mean		4.26

Table 13 clearly indicates that the majority of test-takers believe that an

OPOI principle reduces unnecessary memory load on the part of the test-takers. The results also imply that test-takers believe that listening comprehension test results should be a function of listening skills, rather than unwarranted memory skills. This very problem has been eloquently pointed out by Miller (1999, 106) " … *What many test-takers find most difficult is simply remembering what was said long enough to answer the questions. Each long conversation or lecture lasts between two to three minutes and contains a lot of information. Even many native English speakers cannot remember every word* … "

Table 14. GC: Validity of Being a Speeded Test.

Value	Frequency	Percent
1	15	4.8
2	131	41.7
3	14	4.5
4	116	36.9
5	38	12.1
Mean		3.10

Table 14 shows that only half of the respondents believe that a grammar test should be speeded. This kind of response poses a problem with accepting the degree of validity perceived by the layman test-takers or test-users. In other words, many test-takers do not seem to recognize the importance of speededness to maximize the authenticity of the grammar test. This negative attitude is expected in that no test-takers would appreciate being forced to take a test under the psychological pressure caused by time limitation. Despite these unfavorable responses, however, a desirable grammar test should be administered in a 'speeded' manner in order to ensure more accurate measurement of subconscious acquisition (or implicit/tacit knowledge forming grammatical competence) rather than conscious learning (or explicit knowledge about meticulous grammatical rules). The time constraint makes it possible for the Monitor (Krashen 1985) to be deactivated or suppressed, thus activating subconscious acquisition, which constitutes the basis of genuine communicative skills. The notion of the 'speeded' test is not to be confused with the concept of 'speed' test (e.g. IQ test), which is the opposite of 'power test' in psychometric

terms (Oller 1995).

Table 15. GC: Validity of Formality-specific Assessment.

Value	Frequency	Percent
1	29	9.2
2	85	27.1
3	18	5.7
4	129	41.1
5	53	16.9
Mean		3.29

According to Table 15, a little more than half of the respondents believe that grammar tests should be developed in both a written and a spoken language style. This is presumably due to the fact that many respondents have long been familiarized with the traditional grammar test method which employs a one-sentence format based on written language. It may also be attributed to the fact that the majority of English learners in Korea do not have the awareness of grammar for spoken versus written English, which is an essential concept for effective grammar learning (Brazil 1995).

Table 16. VP: Validity of Being a Speeded Test.

Value	Frequency	Percent
1	16	5.1
2	82	26.1
3	15	4.8
4	136	43.3
5	65	20.7
Mean		3.48

Table 16 reveals that more than half of the respondents believe that a vocabulary test should be speeded, considering the importance of the acquired/ automatized nature of vocabulary in a real-time communication setting (Oller 1995, Widdowson 1989). Some test-takers do not appear to understand the importance of automatized vocabulary power required for good oral skills. It has been pointed out that 'shallow' (i.e. not fully internalized)

vocabulary is often more detrimental to aural comprehension than no vocabulary knowledge (Choi 1997). Therefore, as with the grammar test, it is essential that the vocabulary test be administered in a speeded manner.

Table 17. VP: Validity of Formality-specific Assessment.

Value	Frequency	Percent
1	28	8.9
2	90	28.7
3	15	4.8
4	125	39.8
5	56	17.8
Mean		3.29

Table 17 reveals that, as with the grammar test, a little more than half of the respondents believe that vocabulary tests should be developed in both a written and a spoken language style format. This is presumably because many respondents have long been accustomed to the traditional vocabulary test method which employs no other method than a one-sentence-long written language format. As with the grammar test, it may also be attributed to the fact that the majority of English learners in Korea are not familiar with the essential concepts for effective vocabulary learning, such as the spoken versus written aspects and the productive versus receptive aspects of vocabulary (Mccarthy & Carter 1997, Melka 1997).

Table 18. RC: Importance of Employing Unbiased Topics.

Value	Frequency	Percent
1	6	1.9
2	41	13.1
3	16	5.1
4	182	58.0
5	69	22.0
Mean		3.85

Table 18 illustrates that the test-takers understand that it is important to employ unbiased topics for more valid measurement of reading

comprehension skills. This finding also implies that the respondents are not fully satisfied with the TOEIC reading comprehension test format which contains reading passages whose topics are predominantly based on practical or business English.

Table 19. RC: Importance of Essential Subskill-specific Assessment.

Value	Frequency	Percent
1	6	1.9
2	69	22.0
3	16	5.1
4	175	55.7
5	48	15.3
Mean		3.61

Table 19 indicates that the test-takers feel that it is crucial to measure the basic subskills required for successful reading. They appear to expect the reading comprehension test to measure a variety of fundamental subskills — the ability to identify specific information, to grasp the main idea, to make inferences, and so on. The TOEIC reading test, which is primarily made up of so-called business English, measures the ability to identify specific information rather than to measure the ability to infer or to grasp the main idea.

Table 20. RC: Importance of Speed Reading Caused by OPOI.

Value	Frequency	Percent
1	12	3.8
2	70	22.3
3	28	8.9
4	120	38.2
5	84	26.8
Mean		3.618

Table 20 suggests that many test-takers believe it is important to promote and measure speed reading skills by employing the OPOI principle, which makes it inevitable to present as many reading passages as the total

number of items. The OPOI principle makes it possible for TEPS to present much more authentic linguistic samples than other EFL tests, thus ensuring more accurate and reliable assessment.

It would be insightful to use a text retrieval/concordance program (Wordsmith Tools 3.0, 1999) to compare the simple statistical aspects of reading passages, i.e., the number of words which appear in TEPS and other EFL tests. The total number of words in the reading passages and the vocabulary test in question was 3871 and 970, respectively. Given the fact that the TOEFL includes vocabulary test items within the reading passages, it would be reasonable to compare the total number of words in the TOEFL reading comprehension test with that of the TEPS reading comprehension test and vocabulary test. The amount of reading load in the TEPS reading and vocabulary tests (4841 in this case; ranging from 4500 to 5000 words) is about four times greater than that of the TOEFL reading comprehension test which contains approximately 1000 to 1500 words (four or five passages containing about 250~350 words each: Miller 1999). The ratio is even greater when TEPS is compared with the TOEIC.

Table 21. Desirability of Test Score Report in Specified Areas.

Value	Frequency	Percent
1	1	.3
2	4	1.3
3	38	12.1
4	54	17.2
5	217	69.1
Mean		4.54

Table 21 demonstrates that the majority of test-takers want to receive a score report which provides test scores in categorical areas such as language subskills and subcomponents. This finding strongly suggests that any valid test serves the diagnostic purpose by presenting the test-takers with language skill/component-specific score information. This highlights the fact that valid testing should be geared towards promoting better education.

2. Survey on the First TEPS

The following is the extract from the survey conducted by Korean Gallop

and Marketing Research Korea on the test takers who took the first TEPS test administered on January 31, 1999.

2.1. Summary

The survey was conducted over the phone on the randomly sampled group of 1043 test takers who took the first TEPS in five cities across the country. The margin of error is 95% and the confidence interval ranges from -3.03% to +3.03%. Overall, it shows that the majority of the test takers consider TEPS a valid test of desirable test methods and a potential candidate to substitute for other conventional EFL tests.

2.2. Degree of Discrimination

As Table 22 demonstrates, 65% of all respondents said that TEPS is different from other EFL tests. The discriminatory aspects of TEPS were expressed more by those who had taken TOEIC and the English teachers than by the other average respondents.

Table 22. Perceived Discriminatory Aspects of TEPS.

Group	Number of Respondents	Different	Similar	No Response
Total	1043	65.1	32.7	2.2
Experienced taking TOEIC	618	71.2	28.3	0.5
English teachers for adults	120	77.2	22.0	0.8

2.3. Discriminatory Aspects

As Table 23 indicates, 83 percent of those 680 respondents who believed in the discriminatory features of TEPS had a favorable opinion of TEPS. Among the most favored aspects which discriminates TEPS from the other tests were the valid test methods and the practical English oriented test contents.

Table 23. Major Discriminatory Aspects.

Discriminatory Aspects	%
Valid Test Methods	27.7
Appropriateness of Difficulty Level	20.2
Practical English oriented Test Contents	15.5
High Degree of Discriminatory Power	8.6
Diversity of Content Domains Measured	5.5
Suitability to Korean Students	4.9

The overall survey indicates a strong consensus among the English teachers and test-takers as to the validity of test method facets. As the majority of the respondents have had an opportunity to take other English tests, it can be safely said that the findings from the survey reflect fairly objective opinions comparing the test methods of TEPS with those of other English tests. Considering the social responsibility of a nationally certified test, it is imperative that the judgmental opinions of prospective test-takers regarding the test method facets merit serious consideration and be incorporated in developing a fair and valid measurement tool. In this respect, the current findings from the survey shed some insightful light on developing and validating a general English proficiency test, TEPS.

V. Conclusion

The present research findings from the quantitative and qualitative analyses and the survey strongly endorse the test fairness of TEPS, which is considered an essential condition required for validity (Norton 1997). The survey reveals that the first TEPS has succeeded in meeting the rigorous criterion of test fairness by satisfying the high expectation of the test-takers and test-users. The majority of respondents agreed that TEPS is developed to measure general English proficiency in an accurate and appropriate manner. It is worth noting that most English teaching experts (i.e., the potential test-users) think highly of TEPS, even more than the other EFL tests, especially in terms of the test methods introduced first by TEPS.

The correlational analysis of the SNU students who took TEPS and TOP reveal that TEPS, as an objective test with a multiple-choice format, has a great potential in assessing productive oral skills in an indirect manner. The

case study on the SNU test-takers who took both TEPS and the other EFL tests strongly supports the findings from the correlational analyses. It shows that TEPS has better discriminating power in measuring overall English communicative competence, presumably due to the desirable test methods.

The descriptive statistics also indicate that TEPS measures the overall English proficiency of the target test-taker group in a reliable and valid manner. The dimensionality check shows that the test contents of TEPS prove to be essentially unidimensional, which ensures the appropriate application of IRT to estimating test-takers' ability or latent traits, based on TEPS item responses and analyzing TEPS test results. The comparison between CTT observed scores and IRT true scores suggests that IRT is better equipped than CTT to provide more precise measurement results. It also reveals that serious problems with scoring validity lie with the conventional methods of calculating the total score by adding up the total number of items correct regardless of the complexity of cognitive process involved in solving test items and the salience of language components and subskills to be measured.

The priority research agenda for the future may include a comparability study among TEPS with a model-data fit study. Test equating may not be necessary with the invariance of sample (theta) and parameter (a, b, c) statistics within the ideal framework of IRT (Hambleton & Swaminathan 1985, Hambleton, Swaminathan, & Rogers 1991). With a somewhat inadequate model-data fit, however, the invariance may not be ensured, resulting in the need for test equating research. Employing common or anchor items is essential for test equating to be implemented (Kolen & Brennan 1995). It is virtually impossible, however, to incorporate the concept of anchor test in TEPS, whose items will be easily detected by inquisitive test-takers. Hence, using anchor items for test equating in the testing context in Korea will be viewed simply as a serious threat to test fairness and thus to test validity. Given this testing climate, as a follow-up study, it may be necessary to conduct the comparability study among a series of TEPS tests administered in 1999. In order to maximize the validity of such research, it would be desirable that the identical sample group of test-takers take a series of TEPS tests in question. Such a rigorous research agenda would call for a great deal of logistic and moral support from those who participate in developing and administering TEPS.

References

Bachman, L. F. and Eignor, D. R. (1997) 'Recent advances in quantitative test analysis,' in C. Clapham and D. Corson eds., *Encyclopedia of Language and Education, Vol. 7: Language Testing and Assessment*, 227-242, Netherlands, Kluwer Academic Publishers.

_____ (1996) 'Review of Soul National University Criterion-Referenced English Proficiency Test,' *Language Research* 32.3, 373-383, Language Research Institute, Seoul National University.

_____ and Palmer, A. (1996) *Language Testing in Practice*, Oxford: Oxford University Press.

Bradshaw, J. (1990) 'Test-takers' reactions to a placement test,' *Language Testing* 7, 13-30, London, UK, Edward Arnold.

Brazil, D. (1995) *A Grammar of Speech*, Oxford, UK, Oxford University Press.

Brown, A. (1993) 'The role of test-taker feedback in the test development process: Test takers' reactions to a tape-mediated test of proficiency in spoken Japanese,' *Language Testing* 10, 277-303.

Choi, I-C. (1997) 'Essential Test Method Facets of a General English Proficiency Test and their Validity as Perceived by Test-takers,' *Language Research* 33.4, Language Research Institute, Seoul National University.

Clark, J. L. D. and Swinton, S. S. (1980) *The Test of Spoken English as a Measure of Communicative Ability in English-Medium Instructional Settings*, TOEFL Research Report, No. 7, Princeton, Educational Testing Service.

Hambleton, R. K. and Swaminathan, H. (1985) *Item Response Theory: Principles and Applications*, Boston, MA, Kluwer Publishing.

_____ and Rogers, H. J. (1991) *Fundamentals of Item Response Theory*, Newbury Park, CA, Sage Publications.

Kolen, M. and Brennan, R. (1995) *Test Equating: Methods and Practices*, New York, Springer-Verlag.

Krashen, S. (1985) *Input Hypothesis*, London, Longman Inc.

Lenneberg, E. (1967) *Biological Foundations of Language*, New York, Wiley and Sons.

McCarthy, M. and Carter, R. (1997) 'Written and spoken vocabulary,' in N.

Schmitt and M. McCarthy eds., *Vocabulary: Description, Acquisition and Pedagody*, Cambridge, UK, Cambridge University Press, 20-39.

Melka, F. (1997) 'Receptive vs. productive aspects of vocabulary,' in N. Schmitt and M. McCarthy eds., *Vocabuary: Description, Acquisition and Pedagody*, Cambridge, UK, Cambridge University Press, 84-102.

Miller, G. (1999) *Cracking the TOEFL CBT*, New York, The Princeton Review.

Norton, B. (1997) 'Accountability in language assessment,' in C. Clapham and D. Corson eds., *Encyclopedia of Language and Education, Vol. 7: Language Testing and Assessment*, 313-322, Netherlands, Kluwer Academic Publishers.

Oller, J. W. Jr. (1995) 'Review of Content and Construct Validation of a Criterion-referenced English Proficiency Test by Choi (1994),' *English Teaching*, 50.3, 161-168, The Korean Association of Teachers of English.

Stansfield, C. W. and Kenyon, D. M. (1992) *The Development and Validation of a Simulated Oral Proficiency Interview*, Modern Language Journal, 76, 129-141.

Stout, W., Nandakumar, R., Junker, B., and Chang, H. H. (1991) *Dimtest and Testsim*, Urbana, IL, University of Illinois.

Widdowson, H. G. (1989) 'Knowledg of language and ability for use,' *Applied Linguistics*, 10.2, 128-137.

Wordsmith Tools 3.0. (1999) 'A text retrieval concordance program,' Oxford, UK, Oxford University Press.

최인철 (1998) 'Test of Oral Proficiency (TOP)의 개발 연구,' 어학연구, 34.1, 서울대학교 어학연구소.

Appendix 1.

Table 2. Results of TEPS & Communicative Competence/ Demographic Background (Continued).

Rk	LC	Lv	RC	Lv	GR	Lv	VC	Lv	Tot	Lv	Self-rated Proficiency/Demographic Background/Comment
51	327	1	340	1	68	2	80	2+	815	1	No response.
52	353	1	320	2+	64	2	78	2+	815	1	No response.
53	340	1	310	2+	84	1	80	2+	814	1	No response.
54	327	1	350	1	70	2	66	2	813	1	None; a wide variety of RC passage topics is desirable; LC test content contains practical English; speededness/ time limitation
55	327	1	360	1	58	3+	60	3+	805	1	No response.
56	347	1	320	2+	76	2+	62	2	805	1	No response.
57	313	2+	340	1	76	2+	70	2	799	2+	US; 10(24); TOEFL 603
58	300	2+	360	1	74	2+	64	2	798	2+	No response.
59	320	2+	330	1	74	2+	74	2+	798	2+	None; KATUSA; TOEIC 890; high intermediate proficiency
60	313	2+	330	1	76	2+	78	2+	797	2+	No response.
61	320	2+	350	1	70	2	54	3+	794	2+	No response.
62	307	2+	340	1	76	2+	70	2	793	2+	None; TOEFL 620
63	307	2+	360	1	64	2	60	3+	791	2+	None; Two time exposure in LC is desirable
64	320	2+	330	1	74	2+	66	2	790	2+	Vocabulary test methods are valid; valid methods for measuring genuine communicative competence
65	327	1	310	2+	72	2+	80	2+	789	2+	No response.
66	347	1	320	2+	60	3+	58	3+	785	2+	US; 18(2-3); TOEIC 810
67	340	1	300	2+	68	2	76	2+	784	2+	No response.
68	313	2+	320	2+	68	2	82	1	783	2+	None; TOEFL 600; TOEIC 935; inadequate oral skills
69	327	1	310	2+	74	2+	70	2	781	2+	US; 2(21); two time exposure in LC is desirable
70	327	1	280	2	86	1	86	1	779	2+	None; Content validity of Grammar, Vocabulary, RC; Practical English in LC
71	307	2+	340	1	68	2	62	2	777	2+	No response.
72	273	2	340	1	80	2+	80	2+	773	2+	No response.
73	340	1	280	2	76	2+	76	2+	772	2+	No response.
74	287	2+	340	1	72	2+	72	2+	771	2+	No response.
75	307	2+	310	2+	74	2+	78	2+	769	2+	None; assessment of both Spoken and Written English is valid; a bit boring RC passage topic
76	300	2+	320	2+	68	2	80	2+	768	2+	No response.
77	280	2	340	1	78	2+	68	2	766	2+	No response.
78	327	1	290	2+	70	2	74	2+	761	2+	No response.
79	300	2+	320	2+	74	2+	66	2	760	2+	No response.
80	260	2	330	1	78	2+	90	1	758	2+	None; many test items are difficult
81	313	2+	320	2+	62	2	62	2	757	2+	No response.
82	327	1	290	2+	72	2+	68	2	757	2+	No response.
83	313	2+	310	2+	66	2	68	2	757	2+	No response.
84	333	1	270	2	78	2+	74	2+	755	2+	None; TOEIC 800
85	313	2+	300	2+	70	2	72	2+	755	2+	No response.
86	333	1	310	2+	44	3	66	2	753	2+	No response.
87	253	2	350	1	80	2+	68	2	751	2+	US; 40(6-9)
88	280	2	330	1	64	2	76	2+	750	2+	No response.
89	273	2	340	1	76	2+	60	3+	749	2+	No response.
90	273	2	310	2+	82	1	84	1	749	2+	No response.
91	253	2	340	1	76	2+	78	2+	747	2+	None; TOEIC 830

Rk	LC	Lv	RC	Lv	GR	Lv	VC	Lv	Tot	Lv	Self-rated Proficiency/Demographic Background/Comment
92	227	3+	360	1	86	1	74	2+	747	2+	No response.
93	273	2	330	1	76	2+	68	2	747	2+	US; 1(28); superior to G-TELP; LC is more difficult than TOEIC LC
94	340	1	240	3+	84	1	80	2+	744	2+	No response.
95	280	2	330	1	60	3+	74	2+	744	2+	No response.
96	320	2+	270	2	76	2+	78	2+	744	2+	No response.
97	360	1	260	2	62	2	62	2	744	2+	No response.
98	280	2	310	2+	78	2+	74	2+	742	2+	None; TOEIC 890; intermediate oral skills
99	287	2+	320	2+	60	3+	74	2+	741	2+	No response.
100	320	2+	310	2+	46	3	64	2	740	2+	No response.
101	307	2+	310	2+	58	3+	64	2	739	2+	No response.
102	307	2+	310	2+	62	2	60	3+	739	2+	No response.
103	347	1	270	2	56	3+	64	2	737	2+	None; TOEIC 850; intermediate oral skills; superior to TOEIC
104	300	2+	300	2+	70	2	66	2	736	2+	No response.
105	340	1	270	2	62	2	64	2	736	2+	None; more academically oriented test content
106	273	2	320	2+	76	2+	66	2	735	2+	No response.
107	327	1	270	2	60	3+	78	2+	735	2+	No response.
108	267	2	320	2+	70	2	78	2+	735	2+	No response.
109	300	2+	300	2+	76	2+	58	3+	734	2+	No response.
110	287	2+	320	2+	46	3	80	2+	733	2+	No response.
111	300	2+	300	2+	64	2	68	2	732	2+	No response.
112	233	3+	350	1	72	2+	76	2+	731	2+	No response.
113	253	2	340	1	64	2	72	2+	729	2+	None; TOEIC 740
114	287	2+	290	2+	76	2+	76	2+	729	2+	No response.
115	293	2+	310	2+	64	2	62	2	729	2+	No response.
116	260	2	330	1	66	2	72	2+	728	2+	No response.
117	260	2	340	1	64	2	64	2	728	2+	No response.
118	320	2+	290	2+	56	3+	60	3+	726	2+	No response.
119	313	2+	280	2	66	2	66	2	725	2+	No response.
120	287	2+	310	2+	62	2	66	2	725	2+	No response.
121	347	1	230	3+	70	2	78	2+	725	2+	No response.
122	340	1	260	2	72	2+	52	3+	724	2+	No response.
123	253	2	340	1	68	2	62	2	723	2+	No response.
124	313	2+	300	2+	58	3+	52	3+	723	2+	No response.
125	300	2+	310	2+	52	3+	60	3+	722	2+	No response.
126	333	1	260	2	68	2	60	3+	721	2+	No response.None; academically oriented test content
127	293	2+	280	2	80	2+	68	2	721	2+	No response.None; TOEIC 865; intermediate oral skills;
128	260	2	330	1	66	2	62	2	718	2+	No response.
129	240	3+	340	1	66	2	70	2	716	2+	No response.
130	287	2+	280	2	76	2+	72	2+	715	2+	No response.
131	247	2	330	1	72	2+	66	2	715	2+	No response.
132	320	2+	270	2	60	3+	62	2	712	2+	No response.
133	360	1	230	3+	56	3+	64	2	710	2+	None; TOEIC 885; TOEFL 580; intermediate oral skills; speededness is desirable to maximize discrimination
134	253	2	330	1	60	3+	66	2	709	2+	None; TOEIC 785; more difficult than TOEIC
135	273	2	310	2+	64	2	62	2	709	2+	None; TOEIC 750
136	273	2	300	2+	70	2	66	2	709	2+	No response.
137	307	2+	260	2	70	2	72	2+	709	2+	No response.
138	300	2+	270	2	64	2	72	2+	706	2+	No response.
139	267	2	320	2+	54	3+	64	2	705	2+	None; a wide range of test content
140	307	2+	280	2	54	3+	64	2	705	2+	None; relatively difficult

Rk	LC	Lv	RC	Lv	GR	Lv	VC	Lv	Tot	Lv	Self-rated Proficiency/Demographic Background/Comment
141	320	2+	260	2	66	2	58	3+	704	2+	None; LC and grammar tests desirable; RC test content is more of written/formal English
142	307	2+	280	2	52	3+	64	2	703	2+	No response.
143	267	2	290	2+	68	2	78	2+	703	2+	UK; 2(23)
144	280	2	300	2+	66	2	56	3+	702	2+	None; RC OPOI and various test tasks are desirable
145	280	2	310	2+	48	3	64	2	702	2+	No response.
146	240	3+	330	1	58	3+	72	2+	700	2	No response.
147	300	2+	270	2	64	2	64	2	698	2	None; too easy
148	287	2+	270	2	72	2+	68	2	697	2	No response.
149	280	2	270	2	78	2+	68	2	696	2	No response.
150	300	2+	250	2	70	2	76	2+	696	2	No response.
151	233	3+	330	1	70	2	62	2	695	2	No response.
152	280	2	300	2+	60	3+	54	3+	694	2	No response.
153	347	1	210	3+	74	2+	62	2	693	2	No response.
154	287	2+	280	2	62	2	62	2	691	2	US 2(23) TOEIC 860; inadequate oral skills
155	313	2+	270	2	52	3+	54	3+	689	2	No response.
156	287	2+	280	2	60	3+	62	2	689	2	No response.
157	300	2+	280	2	52	3+	56	3+	688	2	No response.
158	247	2	310	2+	70	2	60	3+	687	2	None; TOEIC 785; straightforward test format
159	180	3	350	1	82	1	74	2+	686	2	No response.
160	240	3+	320	2+	64	2	62	2	686	2	No response.
161	273	2	290	2+	60	3+	62	2	685	2	None; TOEIC 775
162	233	3+	330	1	46	3	74	2+	683	2	None; TOEIC 850; inadequate oral skills
163	247	2	330	1	56	3+	50	3	683	2	No response.
164	280	2	290	2+	58	3+	54	3+	682	2	No response.
165	287	2+	280	2	54	3+	60	3+	681	2	None; RC is too difficult, LC test content is based on fast speech
166	273	2	280	2	52	3+	76	2+	681	2	No response.
167	273	2	290	2+	60	3+	56	3+	679	2	None; test methods are valid; intermission time is required
168	273	2	260	2	76	2+	70	2	679	2	None; measures overall proficiency
169	267	2	290	2+	64	2	56	3+	677	2	No response.
170	267	2	270	2	72	2+	68	2	677	2	No response.
171	293	2+	260	2	66	2	58	3+	677	2	No response.
172	313	2+	240	3+	58	3+	66	2	677	2	No response.
173	267	2	270	2	76	2+	64	2	677	2	No response.
174	260	2	290	2+	66	2	60	3+	676	2	None; TOEIC 730; speededness is desirable
175	260	2	290	2+	70	2	52	3+	672	2	No response.
176	233	3+	310	2+	66	2	62	2	671	2	No response.
177	253	2	290	2+	62	2	66	2	671	2	No response.
178	293	2+	250	2	66	2	60	3+	669	2	No response.
179	287	2+	260	2	70	2	50	3	667	2	No response.
180	233	3+	310	2+	54	3+	70	2	667	2	None; OPOI and speededness are desirable
181	267	2	270	2	66	2	62	2	665	2	No response.
182	227	3+	290	2+	68	2	76	2+	661	2	No response.
183	260	2	260	2	68	2	72	2+	660	2	No response.
184	240	3+	300	2+	56	3+	64	2	660	2	No response.
185	287	2+	250	2	50	3	72	2+	659	2	No response.
186	233	3+	320	2+	58	3+	48	3	659	2	None; much more difficult than TOEIC
187	273	2	240	3+	70	2	74	2+	657	2	No response.
188	273	2	270	2	54	3+	60	3+	657	2	No response.
189	253	2	280	2	60	3+	64	2	657	2	No response.
190	220	3+	300	2+	68	2	68	2	656	2	No response.

Rk	LC	Lv	RC	Lv	GR	Lv	VC	Lv	Tot	Lv	Self-rated Proficiency/Demographic Background/Comment
191	220	3+	310	2+	66	2	60	3+	656	2	No response.
192	247	2	290	2+	60	3+	58	3+	655	2	No response.
193	273	2	270	2	60	3+	52	3+	655	2	No response.
194	207	3+	330	1	60	3+	58	3+	655	2	No response.
195	220	3+	310	2+	64	2	60	3+	654	2	No response.
196	247	2	280	2	64	2	62	2	653	2	No response.
197	267	2	260	2	62	2	62	2	651	2	No response.
198	320	2+	190	3	76	2+	62	2	648	2	No response.
199	260	2	280	2	52	3+	56	3+	648	2	No response.
200	273	2	270	2	54	3+	50	3	647	2	US; 2(20); LC double exposure method is desirable
201	273	2	260	2	60	3+	52	3+	645	2	No response.
202	267	2	260	2	56	3+	58	3+	641	2	No response.
203	193	3	310	2+	68	2	70	2	641	2	No response.
204	287	2+	230	3+	60	3+	64	2	641	2	No response.
205	267	2	260	2	54	3+	58	3+	639	2	No response.
206	180	3	310	2+	72	2+	74	2+	636	2	No response.
207	220	3+	300	2+	64	2	52	3+	636	2	No response.
208	273	2	240	3+	60	3+	62	2	635	2	No response.
209	207	3+	290	2+	62	2	74	2+	633	2	No response.
210	253	2	260	2	56	3+	64	2	633	2	No response.
211	300	2+	210	3+	52	3+	70	2	632	2	US; 3(?) TOEFL 600
212	267	2	260	2	40	4+	64	2	631	2	No response.
213	233	3+	290	2+	56	3+	52	3+	631	2	No response.
214	247	2	260	2	54	3+	68	2	629	2	None; TOEIC 810; inadequate oral skills
215	233	3+	290	2+	58	3+	48	3	629	2	No response.
216	267	2	230	3+	62	2	66	2	625	2	No response.
217	227	3+	270	2	72	2+	56	3+	625	2	None; simple format of LC desirable
218	293	2+	210	3+	56	3+	64	2	623	2	No response.
219	273	2	220	3+	64	2	62	2	619	2	No response.
220	213	3+	270	2	66	2	68	2	617	2	No response.
221	260	2	230	3+	60	3+	66	2	616	2	None; TOEIC 800
222	253	2	250	2	62	2	50	3	615	2	No response.
223	260	2	260	2	46	3	48	3	614	2	None; LC double exposure method is desirable
224	173	3	330	1	48	3	60	3+	611	2	No response.
225	220	3+	290	2+	50	3	50	3	610	2	No response.
226	213	3+	260	2	64	2	68	2	605	2	No response.
227	193	3	280	2	62	2	66	2	601	2	No response.
228	153	4+	330	1	62	2	56	3+	601	2	No response.
229	227	3+	280	2	52	3+	40	4+	599	3+	No response.
230	213	3+	270	2	58	3+	56	3+	597	3+	No response.
231	247	2	230	3+	62	2	58	3+	597	3+	None; LC double exposure method is desirable
232	247	2	230	3+	58	3+	60	3+	595	3+	None; overall test content is valid
233	273	2	240	3+	44	3	38	4+	595	3+	No response.
234	240	3+	240	3+	52	3+	62	2	594	3+	US; 1(?); real-life/ practical English; many RC items
235	173	3	290	2+	64	2	66	2	593	3+	No response.
236	227	3+	260	2	52	3+	48	3	587	3+	No response.
237	227	3+	260	2	50	3	48	3	585	3+	No response.
238	193	3	290	2+	54	3+	48	3	585	3+	None; speededness valid
239	167	3	300	2+	60	3+	56	3+	583	3+	No response.
240	213	3+	290	2+	42	3	36	4+	581	3+	No response.

(The remaining part is omitted as no reponses were made by the respondents.)

* LC: Listening Comprehension; GR: Grammar; VC: Vocabulary; RC: Reading Comprehension; Tot: Total
* Rk: rank; Lv: level

Appendix 2. Eigenvalues from the Factor Analysis

1) Listening Comprehension

Factor	Eigenvalue	Difference
1	19.45595	17.13453
2	2.32142	1.39507
3	.92635	.06575
4	.86060	.08281

2) Grammar

Factor	Eigenvalue	Difference
1	12.74259	10.95009
2	1.79249	.73731
3	1.05518	.26265
4	.79253	.13888

3) Vocabulary

Factor	Eigenvalue	Difference
1	12.93668	11.02994
2	1.90674	.47241
3	1.43433	.68851
4	.74582	.08741

4) Reading Comprehension

Factor	Eigenvalue	Difference
1	9.75580	8.49082
2	1.26497	.40450
3	.86048	.18537
4	.67511	.13742

Department of English Language and Literature
Sungshin Women's University
249-1 Dongsun-dong 3 ga, Sungbook-ku
Seoul 136-742, Korea
E-mail: icchoi@cc.sungshin.ac.kr